Categorias funcionais e de conteúdo na aprendizagem distribucional

Palavras-Chave: aprendizagem distribucional, modelagem computacional, aquisição da linguagem

Autores/as:

MARCOS SÉRGIO ZANCHETTA JUNIOR [IEL-UNICAMP]

PROF. DR. PABLO PICASSO FELICIANO DE FARIA (orientador) [IEL-UNICAMP]

INTRODUÇÃO:

Teorias de aquisição de linguagem tentam descrever e explicar o processo universal pelo qual crianças típicas adquirem sua língua materna, aproximadamente no mesmo período de tempo e sem esforço, não importando que língua seja e nem a diversidade de experiências linguísticas de cada criança (Goodluck, 1991; Mioto et. al, 2007). Uma teoria incluirá, certamente, tanto aspectos universais quanto particulares das línguas e da cognição linguística. Visto que a aquisição da linguagem envolve diversas tarefas – tais como a aquisição fonológica, a lexical (incluindo a morfologia e a semântica das palavras), a sintática etc. – e que para cada um destes níveis linguísticos há variação entre as línguas, imagina-se que tais variações tenham impacto nos mecanismos de aprendizagem, fazendo com que alguns sejam mais eficazes numa língua do que em outra. Diferentes teorias e abordagens debatem sobre quais seriam os mecanismos envolvidos nesse processo, sobre a natureza do conhecimento linguístico e sobre como tais mecanismos interagem com estas propriedades (Tomasello, 1995; Pinker, 2004; entre outros). Esta pesquisa visa contribuir para este debate na medida em que investiga o grau de informatividade das informações distribucionais dos enunciados, extraída a partir das palavras e da ordem relativa entre elas, deixando de lado, neste recorte, outras fontes de informação, assumindo apenas que o aprendiz é capaz de segmentar os enunciados em palavras. Com isso, investigamos a efetividade da aprendizagem distribucional no português (brasileiro).

METODOLOGIA:

O modelo computacional adotado aqui (Faria e Ohashi, 2018; Faria, 2019a, 2019b) implementa um procedimento de três estágios para aprendizagem distribucional das categorias a partir de dados (em

português) de fala dirigida à criança provenientes da base CHILDES (MacWhinney, 2000) e da Coleção "Projeto Aquisição da Linguagem Oral" (CEDAE, UNICAMP):

(i) medir os contextos de distribuição em que cada palavra ocorre;

(ii) comparar o contexto de distribuição para pares de palavras;

(iii) agrupar palavras com distribuições similares.

A primeira etapa envolve medir o contexto em que as palavras ocorrem. Para isso, são coletadas estatísticas de co-ocorrência entre a palavra-alvo e as palavras em seu entorno, armazenando essa informação em uma matriz de co-ocorrência. Nesta, cada linha representa a distribuição de uma dada palavra-alvo e as colunas representam as frequências das palavras contextuais em relação a ela. Assim, a tabela pode ser vista como uma lista de *vetores de contexto* das palavras. Em seguida, na etapa (ii), avalia-se a similaridade entre estes vetores de contexto, que podem ser pensados como pontos em um **espaço multidimensional de possíveis distribuições de palavras. Assim, espera-se que palavras de** mesma categoria sintática tenham distribuições similares, ou seja, estejam relativamente próximas nesse espaço multidimensional. Para calcular a similaridade, utiliza-se o *coeficiente de correlação de postos de Spearman*.

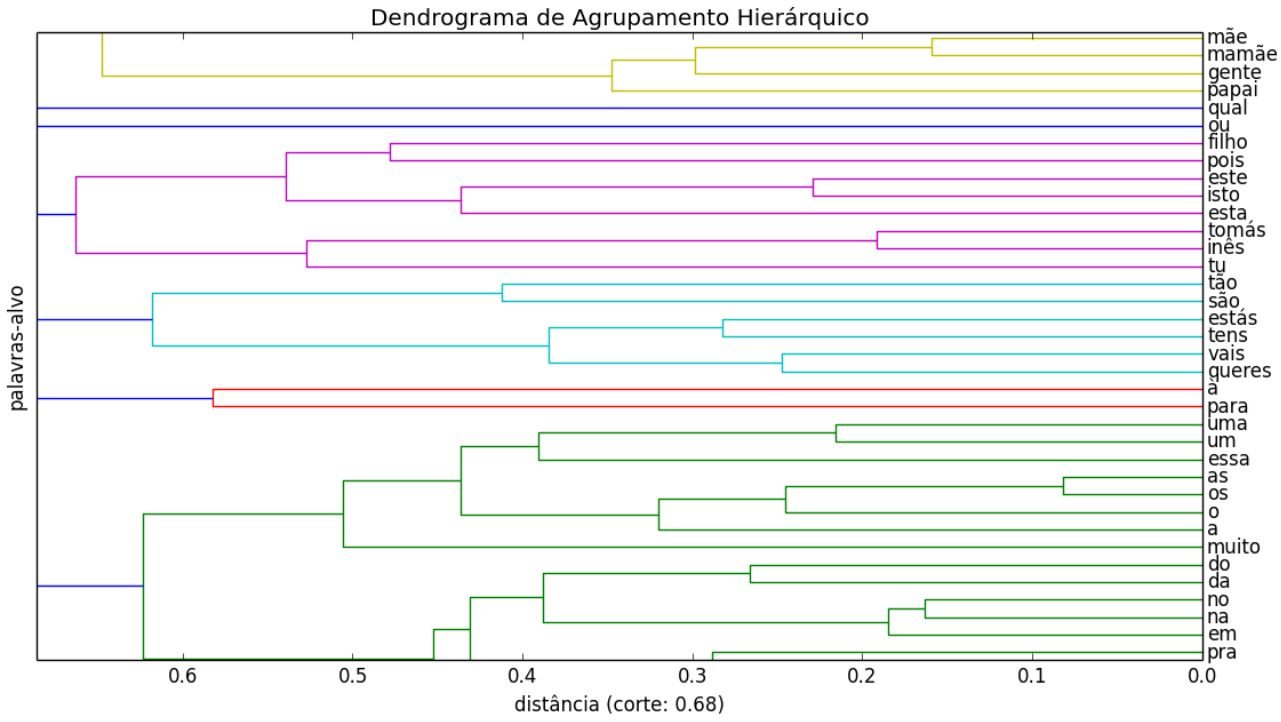

Figura 1. Dendrograma parcial de agrupamentos de palavras.

N a terceira etapa, os agrupamentos são gerados usando o método de *análise de agrupamento hierárquico padrão* (Redington *et al.*, 1998, p. 437). Essencialmente, o algoritmo começa agrupando palavras mais próximas (por similaridade) e segue adiante, agrupando outras palavras entre si ou mesmo com

outros grupos já formados. O método encerra quando um único grupo final é obtido. O agrupamento hierárquico gerado desta forma pode ser representado como um dendrograma, como mostra a Figura 1.

Quatro novas condições experimentais são investigadas nesse estudo: (i) usar apenas palavras funcionais como itens contextuais válidos para classificar as palavras de modo geral e, também, separadamente (os próprios itens funcionais, de um lado, e os de conteúdo, de outro); (ii) usar apenas itens de conteúdo para classificar itens funcionais, itens de conteúdo e as palavras de modo geral e (iii) utilizar as primeiras cem palavras mais comuns no corpus para classificá-lo, divididos em grupos de 10 palavras. Os resultados deste último experimento nos levaram a incluir uma quarta condição experimental, considerando o efeito conjunto das 10 palavras mais frequentes e o grupo das palavras que ocupam o intervalo entre a 41ª palavra mais frequente e a 50ª.

RESULTADOS E DISCUSSÃO:

Na tabela abaixo, mostramos os resultados gerados pelo modelo. A linha "Standard" representa o resultado padrão, sem alterações nos itens contextuais. As linhas "Open" e "Closed" referem-se à classificação de todo o corpus utilizando, como itens contextuais, os itens de conteúdo e os itens funcionais, respectivamente. As quatro últimas linhas apresentam, na primeira posição, o tipo de item contextual e, na segunda, o tipo de palavra que é classificado (p. ex. "Open_closed" representa os resultados em que os itens de conteúdo são usados como contexto e os itens funcionais estão sendo classificados pelo modelo).

Standard	0.63 (p: 0.71 , c: 0.27) | C 0.36 | G 32
Open	0.61 (p: 0.75 , c: 0.20) | C 0.35 | G 66
Closed	0.58 (p: 0.78 , c: 0.15) | C 0.43 | G 69
Open_open	0.67 (p: 0.81 , c: 0.23) | C 0.31 | G 36
Open_closed	0.48 (p: 0.73 , c: 0.10) | C 0.40 | G 39
Closed_open	0.67 (p: 0.75 , c: 0.30) | C 0.35 | G 14
Closed_closed	0.53 (p: 0.69 , c: 0.15) | C 0.43 | G 31

Tabela 1. Resultados do modelo utilizando categorias como itens contextuais

Os resultados apontam que o modelo tem uma melhora de desempenho, em relação ao experimento padrão, quando são classificados os itens de conteúdo, mostrando um F-Score de 0.67 ("Open_open" e "Closed_open"). O resultado referente à classificação das palavras de conteúdo quando utilizamos os itens funcionais como contexto é melhor, no geral, comparando-o com o resultado "Standard", uma vez que, além de um melhor F-Score, há maior precisão e maior cobertura.

A seguir, seguem os resultados do modelo quando são colocadas, como itens contextuais, as palavras de acordo com sua frequência no corpus.

	Novos itens contextuais	100 palavras + comuns | Grupos de 10
1-10	que, é, o, a, não, e, eu, de, tá, olha	0.62 (p: 0.75 , c: 0.21) | C 0.36 | G 49
11-20	aqui, você, vai, pra, cê, um, ah, lá, tem, tu	0.58 (p: 0.64 , c: 0.29) | C 0.65 | G 24
21-30	então,com,hum,uma,está,ela,aí,na,da,ele	0.54 (p: 0.60 , c: 0.24) | C 0.68 | G 23
31-40	isso,mais,agora,no,do,esse,já,mas,né,como	0.43 (p: 0.52 , c: 0.14) | C 0.14 | G 40
41-50	se,quem,foi,bem,vamos,assim,essa,por,ai,fazer	0.39 (p: 0.39 , c: 0.51) | C 0.94 | G 4
51-60	muito,quer,si,mamãe,só,para,me,porque,onde,os	0.34 (p: 0.36 , c: 0.26) | C 0.91 | G 9
61-70	sabe,es,ver,ahn,em,hein,faz,mãe,pois,também	0.33 (p: 0.34 , c: 0.29) | C 0.21 | G 9
71-80	num,vou,pode,depois,conta,deixa,vem,vê,bom,viu	0.34 (p: 0.36 , c: 0.20) | C 0.22 | G 11
81-90	coisa,as,dá,te,pronto,sim,gente,era,casa,põe	0.39 (p: 0.41 , c: 0.28) | C 0.22 | G 9
91-100	sei,quando,meu,cá,mim,isto,tudo,ou,acho,minha	0.30 (p: 0.30 , c: 0.24) | C 0.23 | G 7

Tabela 2 Resultados do modelo utilizando as 100 primeiras palavras mais frequentes como itens contextuais, separadas por grupos de 10

Sobressaem-se, como resultados, a alta precisão do resultado quando consideramos os 10 itens contextuais mais frequentes e a alta cobertura quando consideramos as palavras posicionadas entre as posições 41 e 50.

A seguir, foram realizados novos experimentos considerando como itens contextuais 20 palavras de dois grupos: o grupo das 10 primeiras palavras mais frequentes no corpus e o das 10 palavras que se encontram entre as posições 41 e 50. Esse experimento teve, como objetivo, verificar se alta precisão do primeiro grupo e alta cobertura do segundo influenciariam no resultado do modelo. Além disso, foi feito um novo experimento considerando, como itens, contextuais, a palavra melhor classificada em cada grupo, totalizando 10 itens.

1-10 e 41-50	que,é,o,a,não,e,eu,de,tá,olha,se,quem,foi,bem,vamos, assim,essa,por,ai,fazer	0.61 (p: 0.66 , c: 0.35) | C 0.46 | G 27
Palavra melhor classificada de cada agrupamento	que,aqui,então,isso,se,muito,sabe,num,coisa,sei	0.40 (p: 0.47 , c: 0.14) | C 0.25 | G 65

Tabela 3 Experimentos adicionais com condições variadas de itens contextuais

Os resultados desse experimento apontam que o desempenho do modelo não é afetado positivamente, em relação ao experimento padrão, quando consideramos os dois melhores resultados da tabela 2. Ou seja, ao consideramos o efeito conjunto das 10 palavras mais frequentes, o resultado que apresentou a melhor precisão, e as palavras mais frequentes entre as posições 41 e

50, que apresentou melhor cobertura, o desempenho do modelo mantém-se abaixo do modelo padrão (“Standard”).

BIBLIOGRAFIA

FARIA, P. (2019a). **The role of utterance boundaries and word frequencies in the categorization of words in Brazilian Portuguese through distributional analysis**. *In: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (NAACL’19)*, 152–159.

FARIA, P. (2019b). **Aprendizagem de categorias de palavras por análise distribucional resultados adicionais para Português Brasileiro**. *Diacrítica*, *33*(2), 229-251.

FARIA, P. e OHASHI, G. O. (2018). **A aprendizagem distribucional no português brasileiro: um estudo computacional**. *Revista LinguíStica*, 14(3): 128–156.

GOODLUCK, H. (1991). **Language Acquisition: A Linguistic Introduction**. Blackwell, Oxford.

MACWHINNEY, B. (2000). **The CHILDES Project: Tools for analyzing talk**. Lawrence Erlbaum

Associates, Mahwah, NJ, third edition edition.

MIOTO, C., FIGUEIREDO SILVA, M. C., e LOPES, R. E. V. (2007) **Novo manual de sintaxe**.

Florianópolis: Insular, 3a. ed.

PINKER, S. (2004). **Clarifying the logical problem of language acquisition**. *Journal of Child Language*, 31(4):949–953.

REDINGTON, M., CHATER, N., e FINCH, S. (1998) **Distributional information: A powerful cue for acquiring syntactic categories**. Cognitive science, v. 22, n. 4, p. 425-469.

TOMASELLO, M. (1995). **Language is not an instinct**. Cognitive Development, (10):131–156.